Global Journal on Innovation, Opportunities and Challenges in AAI and Machine Learning
Vol. 5, Issue 1 – 2021
ISSN: 2581-5156

The Role of Machine Learning in Data Analytics: A Review of Unsupervised Learning Algorithms

M. Pavithra[1], P. Divya[1], S. Jayalakshmi[1], P. Manjubala[1]
[1]Department of Computer Science and Engineering,
IFET College of Engineering, Villupuram, Tamilnadu, India.

Abstract

In this chapter, we will discuss unsupervised learning in general terms. An individual engages in unsupervised learning when they have only access to the input data and do not have access to the corresponding output variables. Unsupervised learning has as its goal the manipulation of the underlying structure and distribution of data to better understand it. Because, unlike supervised learning, which was previously discussed on this thread, there are no correct answers and no instructor present, these are referred to as unsupervised learning. To discover and present the interesting structure within the data, algorithms are left to their own devices. Unsupervised learning problems are further divided into two types: clustering problems and association problems. Clustering problems are the most common type of unsupervised learning challenge.

Introduction

In machine learning (ML), knowledge is acquired with the primary goal of generating judgments and recommendations[1]. As depicted in Figure 1, the many learning approaches employed in ML can be generically classified into 4 categories: supervised learning, unsupervised learning, semi-supervised learning, and reinforced learning. A part of supervised learning, classifiers are trained with previously collected data in order to predict or classify previously unobserved instances[2]. A unique feature of Unsupervised learning is that it could obtain the end result from any form of incoming data through it encounters a lack of specific outcome variables or terms. Semi-supervised learning has the ability to build itself from a little portion of labeled data that is later used by the system to segregate the rest of the data, which is highly used in retraining the model after it has been trained[3]. Using rewards and penalties, Reinforcement learning helps in effective interaction with a frequently changing environment to attain specific objectives with all its available possibilities, depending on rewards owned by the system[4]. Without guidance, unsupervised learning is a way of teaching a machine to identify patterns in data that have neither been classed nor labeled and then lets the algorithms conduct the investigation without direction. In this case, the device's objective is to classify unsorted material based on the similarity, structures, and differences, without the benefit of any past training data[5][6][7].

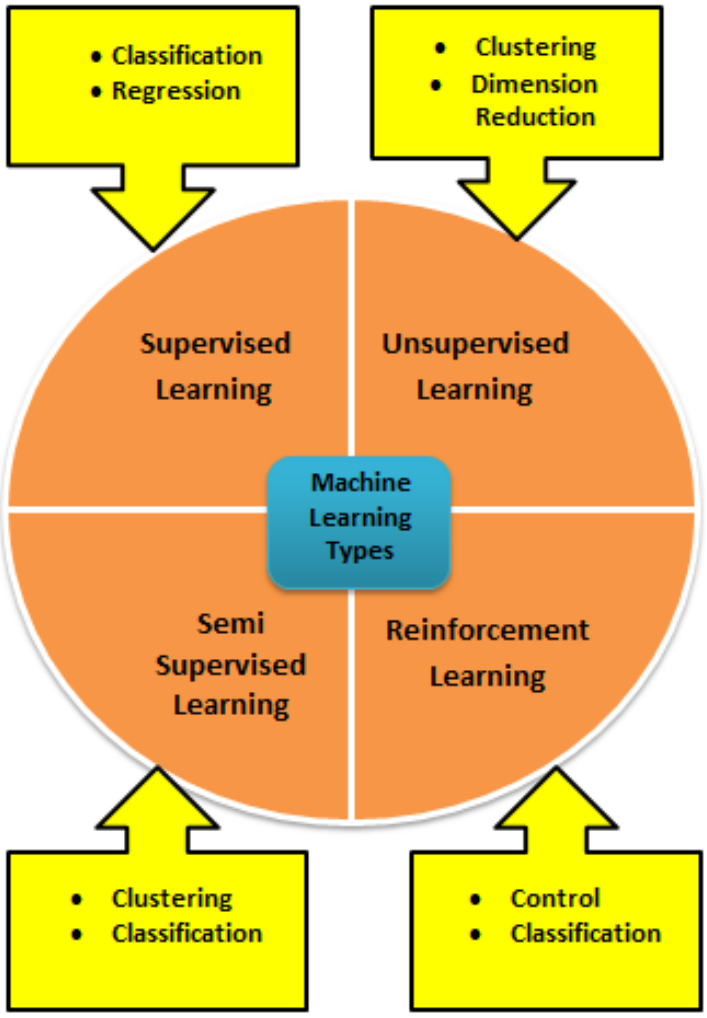

Figure 1.Machine learning Approaches

Without corresponding output labels, an unsupervised study incorporates identifies exciting insights based on the input without using labels. Given that unsupervised learning identifies separate classes without the assistance of an instructor, the precise labeling must be explicitly determined[8-9]. For the most part, unsupervised learning outcomes necessitate user intervention to ensure that the intended classes are correctly identified. Furthermore, although unstructured data is better suitable for practical applications than supervised learning due to the fact that it is more subjective and does not have the explicit purpose of response prediction, its use is expanding all the time [10]. Unsupervised learning is used in various applications, including non-wired digital communication, in the biological sector under DNA classification, building self-learning systems, and others[11,13].

Usage of electrical machines are abundant in recent years, online tracking of those machines without a break is being a challenging task, mainly due to the factors such as weak defective training and test data that sounds high for the improper function of many operating conditions of an electrical machine[12]. Nevertheless, unsupervised learning does provide the possibility of

categorizing and grouping computer reactions over time in a way that can be utilized to detect significant changes in the health of the machine [13]. To be realized, however, Adopting unsupervised learning had paved the way to realize the importance of using fault tolerance methodologies for good product/system outcomes, also that can be implemented without making use of training data. The unsupervised technique of clustering is used in this study to evaluate a prospective predictive maintenance solution [14].

Clustering is essentially a technique for identifying unique groups or classes of observations in a collection of observations. In its most basic form, the similarity between two observations is quantified by the distance between them in the feature space of the observations. In order to quantify the similarity between an instance and a centroid, the Euclidean distance approach is typically utilised[15][16].

After everything is said and done, clustering is a robust tool used in learning and picturing newly arrived data sets. Also, they have been greatly employed in clustering the instances for modeling reasons. It has been represented as a tree structure, with each node corresponding to a separate cluster. Hierarchical clustering can be described as follows: Divisive and agglomerative methods works on split and join nature of clusters that exist which provides much more abstraction of data sets[17]. Several studies have discovered that hierarchical clustering is ideal for small datasets, but partitioning clustering is better suited for large datasets. K-means clustering is one of the most widely used classification techniques nowadays.

Types

Unsupervised Learning Algorithms are classified into the following categories[18]:

It is possible to divide the challenges addressed by the unsupervised learning algorithm into two categories:

Clustering: Clustering is a method of organizing things into clusters [19] so that objects with the most significant amount of similarity remain in one group and have little or no similarity with products in some other group. Cluster analysis have been employed in identifying similarities across data objects and represent those data objects based on the existence of such attributes in the data objects, It is a technique used in data mining.

Association Rules: An association rule is an unsupervised learning strategy [20] that is used to discover correlations between variables in a vast database of data. This function determines a set of items that occur together in a dataset. The business model becomes more effective as a result of the association rule. For example, people who buy a U item (badminton bat) are more likely to buy a V item (shuttle cork) [21].

➢ Clustering can be classified into several categories

Cluster can be classified into two types of subgroups, according to their function[22]:

Hard Clustering: In this clustering, each data will contribute to a group totally or not relate to a group. For illustration, every customer is assigned to one of the ten categories described in the preceding example.

Soft clustering: The term "soft clustering" refers to the process of assigning a likelihood or probability within each data set being in each cluster, rather than placing each piece of data into a splitting criterion[23]. For example, in the situation described above, each customer is allocated a likelihood of being in one of ten clusters of the retail store based on their location.

There are many different types of clustering algorithms.

Because clustering is a subjective endeavour, the number of approaches employed to achieve this goal is numerous. Furthermore, every approach has its own set of principles for determining the degree to which two data points are similar to one another. In reality, there are over 100 clustering methods that have been discovered.

Connection models: This concept is for the closer models in the data space, which makes them greater in similarity with each other than the data points that are away in data space. Two ways can be used with these models. When using the first strategy, they begin by categorizing all data points into independent clusters and then aggregate them when the distance between the data points reduces[24]. The second strategy takes the work of splitting the data points that belongs to a single cluster, and then they have been classified based on the distance calculated between those data points. Furthermore, the selection of the distance function is a matter of personal preference. These models [25] are relatively simple to interpret, but they do not have the scalability to handle large datasets. Models such as the hierarchical clustering algorithm and its modifications are examples of this type.

Centroid models are iterative clustering methods in which the notion of similarity is obtained from the proximity of a data set to the center of the groups, as opposed to the notion of similarity derived from the distance between two data points. In this category, the K-Means [26] clustering algorithm is a standard algorithm that is used in many applications. Because the number of clusters necessary at the end of these models must be specified in advance, it is critical to have foreknowledge about the dataset in question. These models run iteratively in order to locate the local optima [27].

Cluster models based on distribution models: These clustering models are based on the notion of how likely it is that all data sets in a group correspond to the same distribution (For example, Normal, Gaussian)[28]. Regularization is a common problem with these models. An instance of some of those models is the Expectation-Maximization algorithm, which makes use of multidimensional normal distributions to achieve the best possible result.

Density models: As the name suggests, the models under this category are used in discovering the pattern of data points in the given data space. They may be varying in densities of the applied

data points, too, and they are also known as density models. It separates several different concentration zones and assigns the data sets contained in these sections to much the same grouping as the data points outside these regions. DBSCAN and OPTICS are two examples of density models that are widely used.

➢ K- Means Clustering

K-Means learning model is a learning approach that can be used to group data. In contrast to supervised learning, there is no labeled data for this grouping to work. It is performed using K-Means clustering, which involves grouping things into groups that have commonalities to one another but are distinct from the things relating to some other cluster. The letter 'K' represents a numerical value. You must tell the system how many clusters you require it to construct before it can proceed. For example, the number K = 2 denotes the presence of two clusters[29]. There is a method for determining the best or optimum value of K for a given set of information. Let us look at a cricket game to grasp better what k-means are and how they work. Consider the following scenario: you have received data on many cricketers from all across the world, which provides data on the goals allowed by the participant and the wicket made by him in the last ten matches. As a result of this information, we need to divide the data into two groups: the batsmen and the bowlers.

Steps

1. k-means clustering begins with the allocation of two centroids at random (since K=2) [30], which is the initial stage. Centroids are two points that have been given to them. It is significant to mention that the spots can be located everywhere because they are chosen at random. Although they are referred to as centroids, they are not necessarily the center of a given data collection at the outset.
2. The following method calculates the length between all the centroids' data points that were allocated at random. For each point, the range between it and both centroids is determined. The point with the shortest distance is allocated to the centroid with the least distance [31].
3. In order to identify the accurate centroid for these two clusters, the next step must be completed. In order to move the original randomly assigned centroid to the actual centroid of the clusters, the original randomly assigned centroid [32] must be moved.
4. The procedure of determining the distance between two points relocates the centroid. After that, the centroid repositioning comes to an end.

➢ Distance measure by K-means

It follows four types of distance measures

a. Euclidean Distance Measure

The most frequently encountered situation is estimating the difference between the two places. A straight line is a euclidean distance [33] between two points, P and Q, if we have two points, P and Q. It is the distance between two points in Euclidean space measured in meters.

b. Measurement of Manhattan Distance

It is also the addition of the horizontally and vertically components, or the difference between two grid points across planes at odd angles, referred to as the Manhattan distance. It is important to note that we will be calculating the absolute number in order to avoid the effects of negative numbers.

c. The Euclidean Distance Measurement Squared

This measurement is similar to the Euclidean measuring distances, with the exception that this does not include the absolute value at the end [34].

d. Measurement of Cosine Distance

In this scenario, the ratio between both the two lines created by linking the origin point is taken into consideration.

Hierarchical clustering

Hierarchical clustering can be another unsupervised learning approach used to bring together unstructured data points with similar features[35] [36]. For example, hierarchical clustering has been used to combine unprocessed data points that have similar properties. The methods used in hierarchical clustering can be divided into two types.

Every data point is defined as a separate cluster in hierarchical clustering algorithms, which then progressively combine or agglomeration [37] (bottom-up technique) the pairs of clusters that this treatment has formed. Finally, a dendrogram, often known as a tree structure, depicts the hierarchical relationship between the groups.

In contrast, agglomerative hierarchical algorithms [38] treat all pieces of data as if they were part of a single large cluster, with the process of clustering involving the division (Top-down approach) of the single large cluster into numerous small clusters. As a result, agglomerative hierarchical algorithms are used in a variety of applications.

➢ Performing Agglomerative Hierarchical Clustering

Steps to Take we will discuss agglomerative hierarchical clustering, which is the most widely used and essential type of hierarchical clustering. Here are the steps to follow in order to complete the task:

Step 1: Treat each data point as though it were a separate cluster. As a result, we will have, say, K clusters at the start of the game. Thus, in the beginning, the number of data points will also be K.

Step 2: In this step, we will link two data points that are close to one other to build a larger cluster. Thus, the total number of K-1 clusters will be obtained.

Step 3: In order to create more clusters, we must join two closet clusters together. Thus, the total number of K-2 clusters will be reached.

Step 4: To combine all of the data points into a single large cluster, repeat the previous three stages until K = 0, i.e., no more data points to connect.

Step 5: Finally, after creating a single large cluster, dendrograms will be used to divide the cluster into numerous smaller clusters, depending on the nature of the problem.

➢ Distance Measure

The distance between adjacent clusters is easily observed nearest to each other is critical for the hierarchical clustering process. There are several methods for computing the difference across two clusters, and the method used to compute the size determines the grouping rule. These kinds of metrics are referred to as linking methods. Some of the most widely used linkage techniques are listed below:

Single Linkage: This is the smallest link between some of the locations of the groups that are nearest to each other [39].

Complete Linkage: This is the radius points of two separate clusters that can be traveled in one direction. It is among the most widely used linkage strategies because it results in more compact groups than single-linkage.

Average Linkage: When estimating the difference between two clusters, it is necessary to add up the difference between each pairing of datasets, which can then be divided by the larger dataset in order to get the average distance between two clusters. It is also one of the most widely used ways of connection in the world.

Centroid Linkage: In this method of linking, the distance between each cluster's centroid is computed using the centroid distance formula.

Conclusion

Unsupervised learning and supervised learning are often discussed interchangeably. Supervised learning algorithms make use of labeled data. The system takes the collected data and assigns it to specific categories based on whether or not it can predict future outcomes. Manual intervention is typically required with supervised learning algorithms, but they allow for more accurate predictions. In contrast, labeled datasets assist supervised learning algorithms by limiting computational complexity without the need for an extensive training set. Regression and classification are among the most commonly used regression and classification techniques. Task subdivision could be based on the number of examples, categorical attributes, percentage of missing data, and entropy of classes. A lengthy list of data and statistical details are provided on a dataset. To use multiple algorithms, the user must better understand each method's strengths and weaknesses. We might not be able to identify a single classifier that is as effective as a compelling ensemble of classifiers.

References

1. O'Hara, Stephen, Yui Man Lui, and Bruce A. Draper. "Unsupervised learning of human expressions, gestures, and actions." In 2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG), pp. 1-8. IEEE, 2011.
2. S. A. Kumar, H. Kumar, V. Dutt and H. Soni, "Self-Health Analysis with Two Step Histogram based Procedure using Machine Learning," 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), Tirunelveli, India, 2021, pp. 794-799, doi: 10.1109/ICICV50876.2021.9388427.
3. Bouguettaya, Athman, Qi Yu, Xumin Liu, Xiangmin Zhou, and Andy Song. "Efficient agglomerative hierarchical clustering." Expert Systems with Applications 42, no. 5 (2015): 2785-2797.
4. Jahangeer, Gul Shaira Banu, and T. Dhiliphan Rajkumar. "Early detection of breast cancer using hybrid of series network and VGG-16." Multimedia Tools and Applications 80, no. 5 (2021): 7853-7886.
5. S. M. Sasubilli, A. Kumar and V. Dutt, "Improving Health Care by Help of Internet of Things and Bigdata Analytics and Cloud Computing," 2020 International Conference on Advances in Computing and Communication Engineering (ICACCE), Las Vegas, NV, USA, 2020, pp. 1-4, doi: 10.1109/ICACCE49060.2020.9155042.
6. Wu, Ou, Weiming Hu, Stephen J. Maybank, Mingliang Zhu, and Bing Li. "Efficient clustering aggregation based on data fragments." IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 42, no. 3 (2012): 913-926.
7. S.Sasubilli ,A. Kumar,V.Dutt, "Machine Learning Implementation on Medical Domain to Identify Disease Insights using TMS", 2020, Sixth International Conference on Advances in Computing & Communication Engineering Las Vegas USA ICACCE 2020 (22-24 June) ISBN: 978-1-7281-6362-8
8. Zhang, Tian, Raghu Ramakrishnan, and Miron Livny. "BIRCH: an efficient data clustering method for very large databases." ACM sigmod record 25, no. 2 (1996): 103-114.
9. Swarn Avinash Kumar, Harsh Kumar, Srinivasa Rao Swarna, Vishal Dutt, “Early Diagnosis and Prediction of Recurrent Cancer Occurrence in a Patient Using Machine Learning”, European Journal of Molecular & Clinical Medicine, 2020, Volume 7, Issue 7, Pages 6785-6794.
10. Abhishek Kumar, Tvm Sairam, Vishal Dutt, “Machine Learning Implementation For Smart Health Records: A Digital Carry Card”, Global Journal on Innovation, Opportunities and Challenges in AAI and Machine Learning Vol. 3, Issue 1-2019.
11. S. A. Kumar, A. Kumar, V. Dutt and R. Agrawal, "Multi Model Implementation on General Medicine Prediction with Quantum Neural Networks," 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), Tirunelveli, India, 2021, pp. 1391-1395, doi: 10.1109/ICICV50876.2021.9388575.

12. Anuprabhavathi, G., & Rajmohan, R. (2016). Energy-efficient and cost-effective resource provisioning framework for map reduce workloads using dcc algorithm. International Journal of Engineering Science Invention Research & Development, 2(9), 623-628.
13. S. Chandrasekaran and A. Kumar Implementing Medical Data Processing with Ann with Hybrid Approach of Implementation Journal of Advanced Research in Dynamical and Control Systems-JARDCS issue 10, vol.10, page 45-52, ISSN-1943-023X. 2018/09/15.
14. Balaji, S., Balamurugan, B., Kumar, T. A., Rajmohan, R., & Kumar, P. P. (2021). A brief Survey on AI Based Face Mask Detection System for Public Places. Irish Interdisciplinary Journal of Science & Research (IIJSR).
15. Vishal Dutt, Rohit Raturi, Vicente García-Díaz, Sreenivas Sasubilli, "Two-Way Bernoulli distribution for Predicting Dementia with Machine Learning and Deep Learning Methodologies", Solid State Technology, 63(6), pp.: 9528-9546.
16. R. Raturi and A. Kumar " An Analytical Approach for Health Data Analysis and finding the Correlations of attributes using Decision Tree and W-Logistic Modal Process", 2019, IJIRCCE Vol 7, Issue 6, ISSN(Online): 2320-9801 ISSN (Print) : 23209798.
17. Vishvaksenan, K. S., Rajmohan, R., & Kalaiarasan, R. (2017, April). Multi-carrier IDMA system for relay aided cooperative downlink communication with transmitter preprocessing. In 2017 International Conference on Communication and Signal Processing (ICCSP) (pp. 2206-2210). IEEE.
18. Swarn Avinash Kumar, Harsh Kumar, Vishal Dutt, Himanshu Swarnkar, "Role of Machine Learning in Pattern Evaluation of COVID-19 Pandemic: A Study for Attribute Explorations and Correlations Discovery among Variables", (2020): Global Journal on Application of Data Science and Internet of Things, Vol 4 No 2, [ISSN: 2581-4370].
19. Gajalakshmi, R. K., Ananthkumar, T., Manjubala, P., & Rajmohan, R. (2020, July). An Optimized ASM based Routing Algorithm for Cognitive Radio Networks. In 2020 International Conference on System, Computation, Automation and Networking (ICSCAN) (pp. 1-6). IEEE.
20. Balasubramanian, S., Pratheep, S., Rajmohan, R., Kumar, T. A., & Pavithra, M. (2020). SVM Block Based Neural Learning Technique for Identification of Fraudulent Web Pages. Global Journal on Innovation, Opportunities and Challenges in Applied Artificial Intelligence and Machine Learning [ISSN: 2581-5156 (online)], 4(2).
21. Vikas Kumar Singh, Dr. Sanjay Pawar, Lohit Shekam, Vishal Dutt (2020)," Impact Of Covid 19 On Fmcg Sector." Journal of Critical Reviews, 7 (12), 4477-4484. doi:10.31838/ jcr.07.12.640.
22. Cortez, Eli, Altigran S. da Silva, Marcos André Gonçalves, and Edleno S. de Moura. "Ondux: on-demand unsupervised learning for information extraction." In Proceedings of the 2010 ACM SIGMOD International Conference on Management of data, pp. 807-818. 2010.
23. Swarn Avinash Kumar, Harsh Kumar, Vishal Dutt, Himanshu Swarnkar, "Contribution Of Machine Learning Techniques To Detect Disease In-Patients: A Comprehensive Analysisof Classification Techniques", Global Journal on Innovation, Opportunities and Challenges in AAIand Machine Learning, Vol. 3, Issue 1 -2019, ISSN: 2581-5156.

24. Rajkumar, T. Dhiliphan, S. P. Raja, and A. Suruliandi. "Users' click and bookmark based personalization using modified agglomerative clustering for web search engine." International Journal on Artificial Intelligence Tools 26, no. 06 (2017): 1730002.
25. Avants, Brian B., Charles L. Epstein, Murray Grossman, and James C. Gee. "Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain." Medical image analysis 12, no. 1 (2008): 26-41.
26. Swarn Avinash Kumar, Harsh Kumar, Vishal Dutt, Pooja Dixit, "The Role of Machine Learning in COVID-19 in Medical Domain: A Survey", Journal on Recent Innovation in Cloud Computing, Virtualization & Web Applications, Vol 4 No 1 (2020), [ISSN: 2581-544X]
27. Balakrishnan, Guha, Amy Zhao, Mert R. Sabuncu, John Guttag, and Adrian V. Dalca. "An unsupervised learning model for deformable medical image registration." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 9252-9260. 2018.
28. S. Boyapati, S. R. Swarna, V. Dutt and N. Vyas, "Big Data Approach for Medical Data Classification: A Review Study," 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), Thoothukudi, India, 2020, pp. 762-766, doi: 10.1109/ICISS49785.2020.9315870.
29. Davidson, Ian, and S. S. Ravi. "Towards efficient and improved hierarchical clustering with instance and cluster level constraints." State University of New York, Albany, Tech. Rep (2005).
30. Vishal Dutt, Sriramakrishnan Chandrasekaran, Vicente García-Díaz, (2020). "Quantum neural networks for disease treatment identification.", European Journal of Molecular & Clinical Medicine, 7(11), 57-67
31. Nanni, Mirco. "Speeding-up hierarchical agglomerative clustering in presence of expensive metrics." In Pacific-Asia conference on knowledge discovery and data mining, pp. 378-387. Springer, Berlin, Heidelberg, 2005.
32. Suruliandi, A., T. Dhiliphan Rajkumar, and P. Selvaperumal. "Validating The Performance of Personalization Techniques In Search Engine." ICTACT Journal on Soft Computing 5, no. 3 (2015).
33. Swarn Avinash Kumar, Harsh Kumar, Vishal Dutt, Pooja Dixit, "Deep Analysis of COVID-19 Pandemic using Machine Learning Techniques", (2020): Global Journal on Innovation, Opportunities and Challenges in AAI and Machine Learning, Vol 4 No 2, [ISSN: 2581-5156].
34. G. Sasubilli and A. Kumar, "Machine Learning and Big Data Implementation on Health Care data," 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), 2020, pp. 859-864, doi: 10.1109/ICICCS48265.2020.9120906.
35. Stalin, J., R. S. Rajesh, and Smt S. Arun Mozhi Selvi. "A survey on topology and geography based routing protocols in vanets." International Journal of Applied Engineering Research 13, no. 20 (2018): 14813-14822.
36. S. R. Swarna, S. Boyapati, V. Dutt and K. Bajaj, "Deep Learning in Dynamic Modeling of Medical Imaging: A Review Study," 2020 3rd International Conference on Intelligent

Sustainable Systems (ICISS), Thoothukudi, India, 2020, pp. 745-749, doi: 10.1109/ICISS 49785.2020.9315990.

37. Grenager, Trond, Dan Klein, and Christopher D. Manning. "Unsupervised learning of field segmentation models for information extraction." In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05), pp. 371-378. 2005.
38. Swarn Avinash Kumar, Harsh Kumar, Vishal Dutt, Himanshu Swarnkar, "COVID-19 Pandemic analysis using SVM Classifier: Machine Learning in Health Domain", Global Journal on Application of Data Science and Internet of Things, 2020, Vol 4 No. 1.
39. Shalini, S., M. Saravanan, and T. Ananth Kumar. "Design of fault tolerant sequential circuits using selective triple modular redundancy algorithm." International Journal of Pure and Applied Mathematics 119, no. 14 (2018): 1045-1050.

